# Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection

ECCV 2024

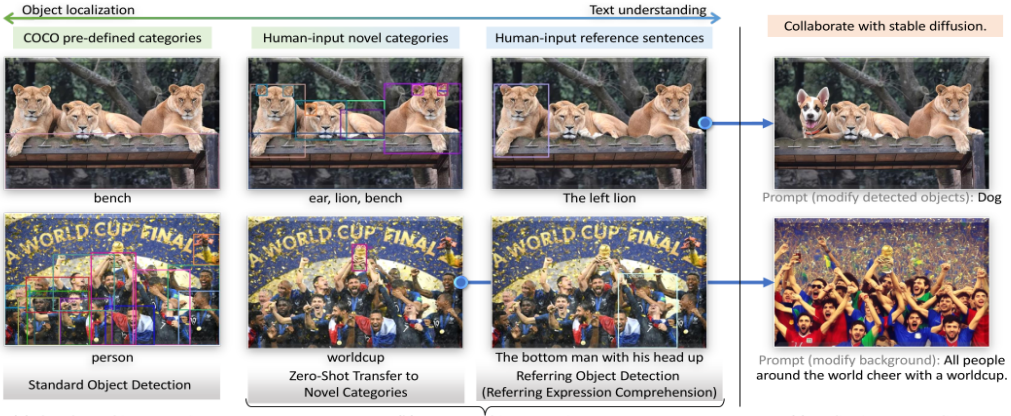
Shilong Liu, Zhaoyang Zeng , Tianhe Ren, Feng Li,Hao Zhang, Jie Yang, Qing Jiang Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang

December 21, 2025

# Objective

## Open-set Object Detection

Detect arbitrary objects specified by human language inputs

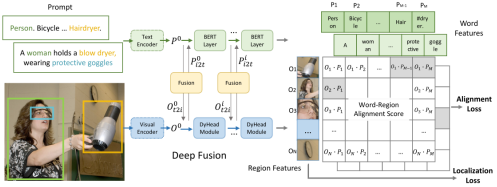


# Novelties

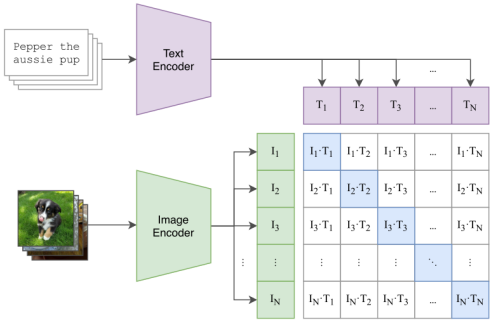
## Novelties

- Large-scale grounded pre-train for zero-shot transfer : CLIP or GLIP
- Tight modality fusion

# CLIP Vs. GLIP

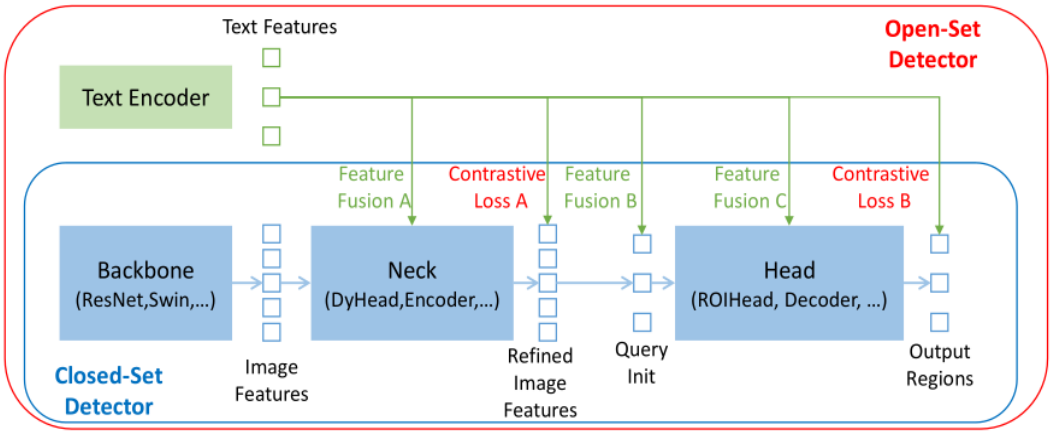



(a) GLIP

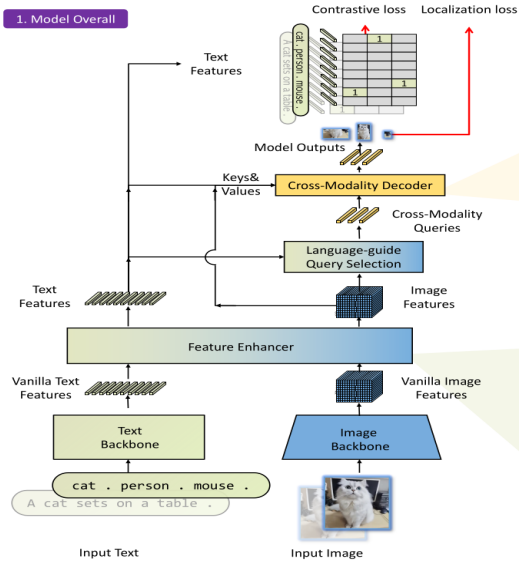



(b) CLIP

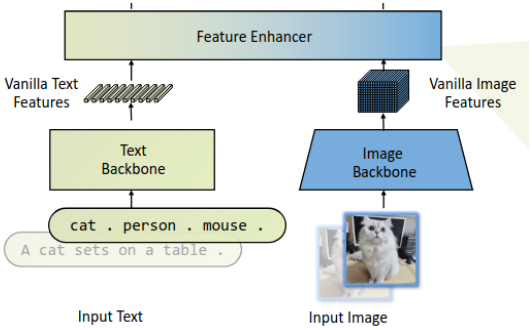
# Tight modality fusion



# GroundingDINO



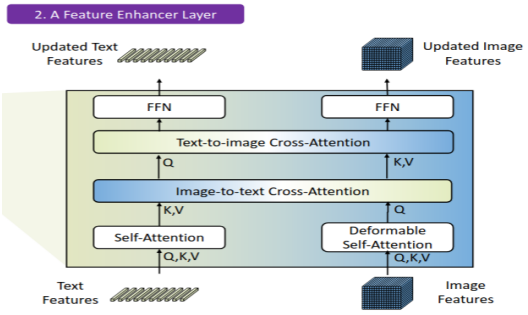
# Feature Extraction and Enhancer




## Feature Extraction

- Image Backbone : Swin Transformer (T, L)
- Text Backbone : BERT

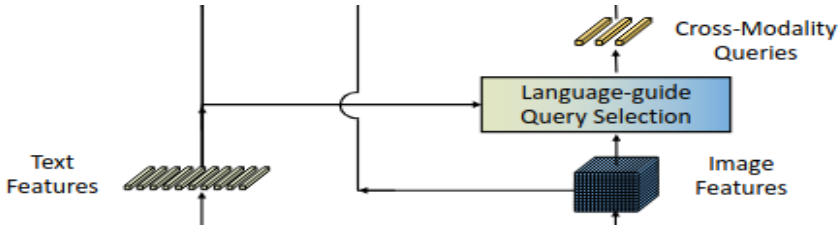
# Feature Extraction and Enhancer




## Feature Enhancer

- Self attention for Text
- Deformable attention for Image
- Image to text attention : Query from Image, Key and value from text
- Text to Image attention : Query from Text, Key and value from image

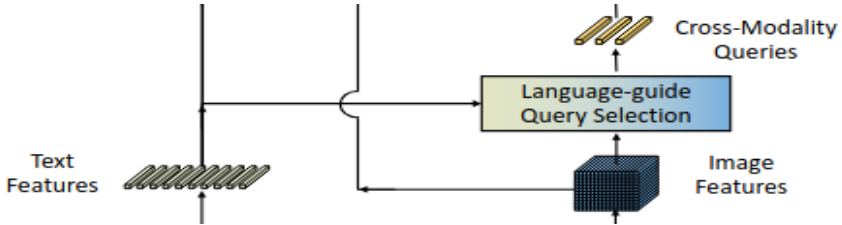
# Language-Guided Query Selection




**To limit number of image tokens**

- $X_I \in \mathbb{R}^{N_I \times d}$ where $N_I \geq 10,000$
- $X_T \in \mathbb{R}^{N_T \times d}$ where $N_T \leq 256$

# Language-Guided Query Selection

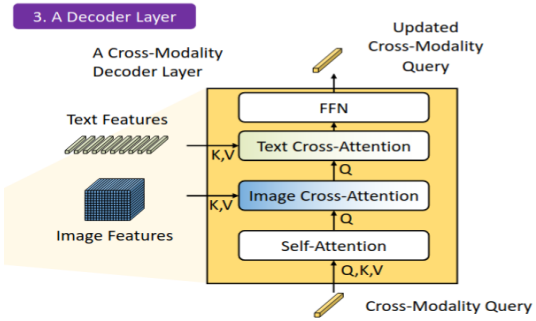



**objective**

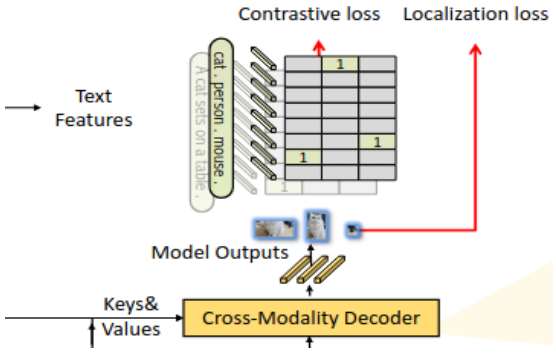
To extract $N_q$ queries from the encoder's image features:

$$I_{N_q} = \text{Top}_{N_q}(\max(X_I X_T^T)), \text{ across the text dimension}$$

# Cross-Modality Decoder



# Loss function




## Loss components

- L1 and GIOU for bounding box regression
- Contrastive loss(focal instead of cross entropy)

# Results

| Model | Backbone | Pre-Training Data | Zero-Shot 2017val | Fine-Tuning 2017val / test-dev |
|---|---|---|---|---|
| Faster R-CNN | RN50-FPN | - | - | 40.2 / - |
| Faster R-CNN | RN101-FPN | - | - | 42.0 / - |
| DyHead-T [5] | Swin-T | - | - | 49.7 / - |
| DyHead-L [5] | Swin-L | - | - | 58.4 / 58.7 |
| DyHead-L [5] | Swin-L | O365,ImageNet21K | - | 60.3 / 60.6 |
| SoftTeacher [50] | Swin-L | O365,SS-COCO | - | 60.7 / 61.3 |
| DINO(Swin-L) [57] | Swin-L | O365 | - | 62.5 / - |
| DyHead-T† [5] | Swin-T | O365 | 43.6 | 53.3 / - |
| GLIP-T (B) [25] | Swin-T | O365 | 44.9 | 53.8 / - |
| GLIP-T (C) [25] | Swin-T | O365,GoldG | 46.7 | 55.1 / - |
| GLIP-L [25] | Swin-L | FourODs,GoldG,Cap24M | 49.8 | 60.8 / 61.0 |
| DINO(Swin-T)† [57] | Swin-T | O365 | 46.2 | 56.9 / - |
| Grounding DINO T (Ours) | Swin-T | O365 | 46.7 | 56.9 / - |
| Grounding DINO T (Ours) | Swin-T | O365,GoldG | 48.1 | 57.1 / - |
| Grounding DINO T (Ours) | Swin-T | O365,GoldG,Cap4M | 48.4 | 57.2 / - |
| Grounding DINO L (Ours) | Swin-L | O365,OI [19],GoldG | **52.5** | **62.6** / **62.7** (**63.0** / **63.0**)* |
| Grounding DINO L (Ours) | Swin-L | O365,OI,GoldG,Cap4M,COCO,RefC | **60.7**‡ | **62.6** / - |

**Table 2:** Zero-shot domain transfer and fine-tuning on COCO. * The results in brackets are trained with 1.5× image sizes, i.e., with a maximum image size of 2000. †The models map a subset of O365 categories to COCO for zero-shot evaluations. ‡This is not a real zero-shot performance as we add COCO data during model training. We list the result here for a reference.

# Results

| Model | Backbone | Pre-Training Data | MiniVal [18] AP | MiniVal [18] APr/APc/APf |
|---|---|---|---|---|
| *Zero-Shot Setting* | | | | |
| GLIP-T (C) | Swin-T | O365,GoldG | 24.9 | 17.7/19.5/31.0 |
| GLIP-T | Swin-T | O365,GoldG,Cap4M | 26.0 | 20.8/21.4/31.0 |
| DetCLIPv2 | Swin-T | O365,GoldG,CC15M | 40.4 | 36.0/41.7/40.0 |
| Grounding DINO T | Swin-T | O365,GoldG | 25.6 | 14.4/19.6/32.2 |
| Grounding DINO T | Swin-T | O365,GoldG,Cap4M | 27.4 | 18.1/23.3/32.7 |
| Grounding DINO L | Swin-L | O365,OI,GoldG,Cap4M, COCO,RefC | 33.9 | 22.2/30.7/38.8 |
| *Fine-Tune Setting* | | | | |
| MDETR | RN101 | GoldG,RefC | 24.2 | 20.9/24.9/24.3 |
| Mask R-CNN | RN101 | - | 33.3 | 26.3/34.0/33.9 |
| DetCLIPv2 [51] | Swin-T | O365,GoldG,CC15M | 50.7 | 44.3/52.4/50.3 |
| Grounding DINO T | Swin-T | O365,GoldG | **52.1** | 35.4/51.3/55.7 |

**Table 3:** Model results on LVIS.

# Results

| Model | Language Input | Backbone | Model Size | Pre-Training Data | Test $AP_{average}$ | Test $AP_{median}$ |
|---|---|---|---|---|---|---|
| *Zero-Shot Setting* | | | | | | |
| MDETR [18] | ✓ | ENB5 [46] | 169M | GoldG,RefC | 10.7 | 3.0 |
| OWL-ViT [35] | ✓ | ViT L/14(CLIP) | >1243M | O365, VG | 18.8 | 9.8 |
| GLIP-T [25] | ✓ | Swin-T | 232M | O365,GoldG,Cap4M | 19.6 | 5.1 |
| OmDet [60] | ✓ | ConvNeXt-B | 230M | COCO,O365,LVIS,PhraseCut | 19.7 | 10.8 |
| GLIPv2-T [59] | ✓ | Swin-T | 232M | O365,GoldG,Cap4M | 22.3 | 8.9 |
| DetCLIP [52] | ✓ | Swin-L | 267M | O365,GoldG,YFCC1M | 24.9 | 18.3 |
| Florence [54] | ✓ | CoSwinH | ≈841M | FLD900M,O365,GoldG | 25.8 | 14.3 |
| Grounding DINO T(Ours) | ✓ | Swin-T | 172M | O365,GoldG | 20.0 | 9.5 |
| Grounding DINO T(Ours) | ✓ | Swin-T | 172M | O365,GoldG,Cap4M | 22.3 | 11.9 |
| Grounding DINO L(Ours) | ✓ | Swin-L | 341M | O365,OI,GoldG,Cap4M,COCO,RefC | **26.1** | **18.4** |
| *Few-Shot Setting* | | | | | | |
| DyHead-T [5] | ✗ | Swin-T | ≈100M | O365 | 37.5 | 36.7 |
| GLIP-T [25] | ✓ | Swin-T | 232M | O365,GoldG,Cap4M | 38.9 | 33.7 |
| DINO-Swin-T [57] | ✗ | Swin-T | 49M | O365 | 41.2 | 41.1 |
| OmDet [60] | ✓ | ConvNeXt-B | 230M | COCO,O365,LVIS,PhraseCut | 42.4 | 41.7 |
| Grounding DINO T(Ours) | ✓ | Swin-T | 172M | O365,GoldG | **46.4** | **51.1** |
| *Full-Shot Setting* | | | | | | |
| GLIP-T [25] | ✓ | Swin-T | 232M | O365,GoldG,Cap4M | 62.6 | 62.1 |
| DyHead-T [5] | ✗ | Swin-T | ≈100M | O365 | 63.2 | 64.9 |
| DINO-Swin-T [57] | ✗ | Swin-T | 49M | O365 | 66.7 | 68.5 |
| OmDet [60] | ✓ | ConvNeXt-B | 230M | COCO,O365,LVIS,PhraseCut | 67.1 | 71.2 |
| DINO-Swin-L [57] | ✗ | Swin-L | 218M | O365 | 68.8 | 70.7 |
| Grounding DINO T(Ours) | ✓ | Swin-T | 172M | O365,GoldG | **70.7** | **76.2** |

**Table 4:** Model results on the ODinW benchmark.

# Results

| Method | Backbone | Pre-Training Data | Fine-tuning | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | val | testA | testB | val | testA | testB | val | test |
| MAttNet [53] | R101 | None | ✓ | 76.65 | 81.14 | 69.99 | 65.33 | 71.62 | 56.02 | 66.58 | 67.27 |
| VGTR [10] | R101 | None | ✓ | 79.20 | 82.32 | 73.78 | 63.91 | 70.09 | 56.51 | 65.73 | 67.23 |
| TransVG [7] | R101 | None | ✓ | 81.02 | 82.72 | 78.35 | 64.82 | 70.70 | 56.94 | 68.67 | 67.73 |
| VILLA_L* [11] | R101 | CC, SBU, COCO, VG | ✓ | 82.39 | 87.48 | 74.84 | 76.17 | 81.54 | 66.84 | 76.18 | 76.71 |
| RefTR [26] | R101 | VG | ✓ | 85.65 | 88.73 | 81.16 | 77.55 | 82.26 | 68.99 | 79.25 | 80.01 |
| MDETR [18] | R101 | GoldG,RefC | ✓ | 86.75 | 89.58 | 81.41 | 79.52 | 84.09 | 70.62 | 81.64 | 80.89 |
| DQ-DETR [45] | R101 | GoldG,RefC | ✓ | 88.63 | 91.04 | 83.51 | 81.66 | 86.15 | 73.21 | 82.76 | 83.44 |
| GLIP-T(B) | Swin-T | O365,GoldG | | 49.96 | 54.69 | 43.06 | 49.01 | 53.44 | 43.42 | 65.58 | 66.08 |
| GLIP-T | Swin-T | O365,GoldG,Cap4M | | 50.42 | 54.30 | 43.83 | 49.50 | 52.78 | 44.59 | 66.09 | 66.89 |
| Grounding DINO T (Ours) | Swin-T | O365,GoldG | | 50.41 | 57.24 | 43.21 | 51.40 | 57.59 | 45.81 | 67.46 | 67.13 |
| Grounding DINO T (Ours) | Swin-T | O365,GoldG,RefC | | 73.98 | 74.88 | 59.29 | 66.81 | 69.91 | 56.09 | 71.06 | 72.07 |
| Grounding DINO T (Ours) | Swin-T | O365,GoldG,RefC | ✓ | 89.19 | 91.86 | 85.99 | 81.09 | 87.40 | 74.71 | 84.15 | 84.94 |
| Grounding DINO L (Ours)* | Swin-L | O365,OI,GoldG,Cap4M,COCO,RefC | ✓ | **90.56** | **93.19** | **88.24** | **82.75** | **88.95** | **75.92** | **86.13** | **87.02** |

**Table 5:** Top-1 accuracy comparison on the referring expression comprehension task. We mark the best results in bold. All models are trained with a ResNet-101 backbone. We use the notations "CC", "SBU", "VG", "OI", "O365", and "YFCC" for Conceptual Captions [44], SBU Captions [36], Visual Genome [20], OpenImage [22], Objects365 [63], YFCC100M [47] respectively. The term "RefC" is used for RefCOCO, RefCOCO+, and RefCOCOg three datasets. * There might be a data leak since COCO includes validation images in RefC. But the annotations of the two datasets are different.